Procesamiento del Lenguaje Natural, Revista nº 65, septiembre de 2020, pp. 93-96 recibido 30-04-2020 revisado 26-05-2020 aceptado 10-06-2020

COST Action "European network for Web-centred linguistic data science" (NexusLinguarum)

Acción COST "Red europea para la ciencia de datos lingüísticos centrada en la web" (NexusLinguarum)

Thierry Declerck[1], Jorge Gracia[2], John P. McCrae[3]
[1]DFKI GmbH, Multilinguality and Language Technology Lab
[2]Aragon Institute of Engineering Research, University of Zaragoza
[3]Insight SFI Research Centre for Data Analytics, National University of Ireland Galway
declerck@dfki.de, jogracia@unizar.es, john.mccrae@insight-centre.org

Abstract: We present the current state of the large "European network for Web-centred linguistic data science". In its first phase, the network has put in place several working groups to deal with specific topics. The network also already implemented a first round of Short Term Scientific Missions (STSM)
Keywords: linguistic data science, multilingualism, linguistic linked data, language resources

Resumen: Presentamos el estado actual de la "Red Europea para la ciencia de datos lingüísticos centrada en la Web". En su primera fase, el projecto ha establecido varios grupos de trabajo para tratar temas específicos. La red también implementó una primera ronda de Misiones Científicas de Corto Plazo (la sigla STSM en Ingles, para Short Term Scientifc Mission).
Palabras clave: ciencia de datos lingüísticos, multilingüismo, datos lingüísticos anlazadoss, recursos lingüísticos

1 Introduction

We report on the current state of development of the "European network for Web-centred linguistic data science" (NexusLinguarum), which is a recently started COST Action.[1]

The main aim of NexusLinguarum is to promote synergies between linguists, computer scientists, terminologists, and other stakeholders in industry and society, in order to investigate and extend the area of linguistic data science. The network understands linguistic data science as a subfield of the expanding "data science", which focuses on the systematic analysis and study of the structure and properties of data at a large scale, along with methods and techniques to extract new knowledge and insights from it. Linguistic data science is a specific case, which is concerned with providing a formal basis to the analysis, representation, integration, and exploitation of language data (syntax, morphology, lexicon, etc.). NexusLinguarum brings thus the topic of linguistic data into this big data context.

In order to support the study of linguistic data science, the network is aiming at the construction at Web scale of a mature holistic ecosystem of multilingual and semantically interoperable linguistic data. Such an ecosystem is needed to foster the systematic cross-lingual discovery, exploration, exploitation, extension, curation, and quality control of linguistic data. For this, NexusLinguarum is investigating the combination of linked data (LD) technologies, natural language processing (NLP) techniques and multilingual language resources (LRs) (bilingual dictionaries, multilingual corpora, terminologies, etc.). This combination seems to offer the potential to enable such an ecosystem that will allow for transparent information flow across linguistic data sources in multiple languages, by addressing in a principled way the semantic interoperability problem.

This combination builds on and further extends the recent development of the so-called

[1]See https://nexuslinguarum.eu/. The action started in October 2019, for a duration of 4 years. "COST" stands for "European Cooperation in Science and Technology". See https://www.cost.eu/.

ISSN 1135-5948. DOI 10.26342/2020-65-11

Linguistic Linked Open Data cloud (LLOD).[2] LLOD grounds on linked data to share and interlink linguistically relevant data sources. Specifically, linked data refers to the recommended best practices for exposing, sharing, and connecting structured data on the Web[3] and builds on Semantic Web recommendations and World Wide Web Consortium (W3C) standards such as Resource Description Framework (RDF), RDF Schema (RDFS), and Web Ontology Language (OWL). In this, NexusLinguarum closely cooperates with on-going H2020 projects dealing with LLOD topics, Elexis[4] and Prêt-à-LLOD.[5]

2 Structure of the Network

The network counts on a very large number of participants, with institutions from 37 so-called COST Member Countries, 3 institutions from Near Neighbour Countries (Belarus, Georgia, Kosovo), 2 International Partner Countries (United States, Singapore) and 1 Specific Organisation (the Translation Centre for the Bodies of the European Union).

NexusLinguarum is co-ordinated by the University of Zaragoza, supported by the Polytechnic University of Madrid for administrative and financial matters. The chair of the Action is Jorge Gracia (University of Zaragoza, Spain), the Vice Chair is John McCrae (National University of Ireland, Galway) and the Scientific Communication Manage is Thierry Declerck (German Research Center for Artificial Intelligence). The Grant Holder Scientific Representative for the network is Elena Montiel (Polytechnic University of Madrid). The co-ordinator for the Short Term Scientific Missions is Penny Labropoulou (Athena Research Center) and Vojtech Svatek (University of Economics, Prague) is the ITC conference manager.

At its kick-off meeting, the network structured itself in 5 working groups, which are briefly described in the following sections.

2.1 Working Group 1 - Linked data-based language resources (leader: Milan Dojchinovski, Czech Technical University, Prague and Julia Bosque-Gil, University of Zaragoza)

WG 1 is laying the foundations and is developing best practices for the evolution, creation, improvement, diagnosis, repair, and enrichment of LLOD resources and value chains.

In the first phase of the action, this WG follows the task of identifying the current state of the LLOD cloud, with its problems and challenges, as well as to identify the state of the art and challenges in the current linguistic data models, especially with respect to under-resourced languages and domains that could and should be integrated in the LLOD.

One of the outcomes of WG1 will result in a matrix of datasets and languages, which can be shared with related work done in the Elexis project, which is aiming at developing a Matrix Dictionary in the field of eLexicography.

2.2 Working Group 2 - Linked data-aware NLP services (leader: Marieke van Erp, KNAW Humanities Cluster, Amsterdam)

This WG focuses on the application of linguistic data science methods including linked data to enrich NLP tasks in order to take advantage of the growing amount of linguistic (open) data available on the Web.

A main task in the first year of the action consists in identifying a list of existing standards, datasets annotations, NLP services and experts to connect to the COST action.

Additional tasks to be supported by this WG are for example the detection and description of LLOD data sets that can support a series of NLP tasks, like Natural Language Generation or Machine Translation.

2.3 Working Group 3 - Support for linguistic data science (leader: Dagmar Gromann, Centre for Translation Studies, Vienna, and Amaryllis Mavragani, University of Stirling, UK)

This working group focusses in the first year on fostering the study of linguistic data by following data analytic techniques at a large scale in combination with LLOD and linked

[2] http://linguistic-lod.org/llod-cloud

[3] https://www.w3.org/DesignIssues/LinkedData.html

[4] https://elex.is/

[5] https://www.pret-a-llod.eu/

data-aware NLP techniques. Big data and linguistic information. In this task, big data sources and state-of-the-art statistical analysis will be studied in combination to LLOD in order to better understand the language.

Visual analytics will be also considered for this task. This will have an impact on all sub-domains of linguistics, from typology to syntax to comparative linguistics.

WG3 is touching issues related to big data and linguistic information, as well as deep learning and neural approaches for linguistic data, in a multilingual setting.

2.4 Working Group 4 - Use cases and applications (Leader: Sara Carvahlo, University of Aveiro, Portugal, and Ilan Kernerman, K Dictionaries, Israel)

WG 4 is dealing with the identification of use cases that will demonstrate the possible deployment of LLOD related technologies. Use cases identified are in the legal domain, as well as in the digital humanities and social sciences, in order to show how linguistic data science can deeply influence studies in those fields.

At the first meeting of the WG, there was an agreement that setting up use cases that are broader in scope, particularly at this initial stage, could help get things in motion. Currently, five main topics are considered, although others will be added in later stages: legal domain, humanities and social sciences, linguistics, life science, and technology.

2.5 **Working Group 5 - Management and dissemination** (leader: Jorge Gracia, University of Zaragoza)

This WG deals with the day-to-day management and administrative coordination, as well as cross-WG and external communication and capacity building. Recently, a new task was added for scientific communication strategy. While this WG seems to be classical management one, it has to deal with the fact that the COST Actions are funding networking activities and not directly research work.

Therefore a main challenge consists in establishing cooperative relations with Research and Development programmes ant also to identify the most relevant Short Term Scientific Missions that can contribute to this kind of cooperation.

3 Possible Impacts of the Network

NexusLinguarun contributes to knowledge creation and transfer in several aspects.

Firstly, through training programs, scientific/industry events, datathons[6], which serve to promote and teach linguistic data science and its related technologies to people from both academia and industry.

Secondly, NexusLinguarum is contributing to a series of W3C standardisation activities in the field of Language Resources.

Thirdly, the project is aiming at disseminating its topics to a broader public, including popular science publications, blog posts on the Web, and contributions to crowdsourced resources, like Wikipedia.

Finally, the Action is committed to the design of a common curriculum for a Europe-wide master degree in linguistic data science, as a means to ensure knowledge transfer to a new generation of researchers and practitioners coming from different disciplines and with different backgrounds, in the topics related to linguistic data science.

4 Conclusion

We briefly presented the current state of development of the European network for Web-centred linguistic data science” (NexusLinguarum), which we think could greatly benefit to be discussed at the SEPLN conference, as the network could get impulses from the SEPLN communities, which are dealing with large varieties of languages.

Acknowledgments

Work presented here was supported in part by the COST Action CA18209 - NexusLinguarum “European network for Web-centred linguistic data science”, the project Prêt-à-LLOD, under grant agreement no. 825182, and the ELEXIS project, under grant agreement no. 731015.

[6] NexusLinguarum will for example organise the next SD-LLOD Datathon and the next edition of the Language, Data and Knowledge (LDK 2021) conference. See http://2019.ldk-conf.org/sd-llod-2019/ for the past edition of those events.

References

Bellandi, A., E. Giovannetti, S. Piccini, and A. Weingart. 2017. Developing LexO: a collaborative editor of multilingual lexica and termino-ontological resources in the humanities. In *Proceedings of Language, Ontology, Terminology and Knowledge Structures Workshop (LOTKS 2017)*, Montpellier, France, September. Association for Computational Linguistics.

Berners-Lee, T. 2006. Design issues: Linked data. http://www.w3.org/DesignIssues/ LinkedData.html (August 24, 2020).

Chiarcos, C., J. McCrae, P. Cimiano, and C. Fellbaum. 2013. Towards open data for linguistics: Linguistic linked data. In A. Oltramari, et al., editors, *New Trends of Research in Ontologies and Lexical Resources.* Springer, Heidelberg, Heidelberg, Germany.

Cimiano, P., C. Chiarcos, J. P. McCrae, and J. Gracia. 2020. *Linguistic Linked Data - Representation, Generation and Applications*. Springer.

McCrae, J., G. A. de Cea, P. Buitelaar, P. Cimiano, T. Declerck, A. Gómez-Pérez, J. Gracia, J., L. Hollink, E. Montiel-Ponsoda, D. Spohr, and T. Wunner. 2012. Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46(6):701–719.

Bosque-Gil, J., D. Lonke, J. Gracia, and I. Kernerman. 2019. Validating the OntoLex-lemon lexicography module with K Dictionaries' multilingual data. In *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference*, pages 726–746, Brno, Czech Republic, October. Lexical Computing CZ s.r.o.